A Holistic Visual Place Recognition Approach Using Lightweight CNNs for Significant ViewPoint and Appearance Changes

Ahmad Khaliq, Shoaib Ehsan, Zetao Chen, Michael Milford, and Klaus McDonald-Maier

***Abstract*—This article presents a lightweight visual place recognition approach, capable of achieving high performance with low computational cost, and feasible for mobile robotics under significant viewpoint and appearance changes. Results on several benchmark datasets confirm an average boost of 13% in accuracy, and 12*x* average speedup relative to state-of-the-art methods.**

***Index Terms*—Convolutional neural network (CNN), feature encoding, robot localization, vector of locally aggregated descriptors (VLADs), visual place recognition (VPR).**

I. Introduction

Given a query image, an image retrieval system aims to retrieve all images from a large database that contain similar objects as in the query image. Visual place recognition (VPR) can also be interpreted as an image retrieval system that tries to recognize a place by matching it with the places from the stored database [1]. A place database is a simplest way to represent a particular environment where appearance based information is stored as an image with no pose related data. Other VPR techniques use topological maps which contain relative information about the places in an environment (can be an ordered collection of images) and metric maps which are even more accurate in terms of absolute scale of the environment (such as distance and landmark position) but difficult to build and maintain. Two image matching techniques; single image and sequence of images are employed by the VPR community. This article focuses on database-centric place remembering approach coupled with single image matching, thus, place recognition is solely based on appearance similarity and image retrieval techniques are applicable [2].

As with a range of other computer vision applications, convolutional neural networks (CNNs) have shown promising results for VPR and managed to shift the focus from traditional hand-crafted feature descriptors [3], [4] to CNNs [5]–[7]. Using a pretrained CNN for VPR, there are three standard approaches to produce a compact image representation: (a) the entire image is directly fed into the CNN and responses from convolutional layers are extracted [5]; (b) CNN is applied on user-defined regions of the image and prominent activations are pooled from the layers representing those regions [6]; (c) the entire image is fed into the CNN and salient regions are identified by directly extracting distinguishing patterns based on convolutional layers responses [7], [8]. Generally, global image representations retrieved from category (a) are not robust against strong viewpoint variations and partial occlusion. Image representations emerging from category (b) usually handle viewpoint changes better but are computationally intensive. Image representations resulting from category (c) address both the appearance and viewpoint variations. In this article, we focus on category (c).

The work done in [7] and [8] are considered state-of-the-art in identifying prominent regions by directly extracting unique patterns based on convolutional layers' responses. Chen *et al.* [7], used VGG-16 network [9] pretrained on ImageNet [10] and used late convolutional layers for regions identification. For regions-based feature encoding, $10k$ bag-of-words (BoW) [11] codebook is employed. The system is tested on five benchmark datasets with AUC-PR curves [12] as the evaluation metric. It claims to outperform FABMAP [13], SEQSLAM [14], and other image retrieval pooling techniques including Cross-Pool [15], Sum/Average-Pool [16], and Max-Pool [17].

Despite its good AUC-PR performance, the method proposed in [7] has some shortcomings. A common strategy for improving CNN accuracy is to make it deeper by adding more layers (provided sufficient data and strong regularization). However, increasing network's size results in increased computation and using more memory both at time of training and testing (such as, for storing outputs of intermediate layers and for storing parameters) is not ideal for resource-constrained robots that are usually battery-operated. Using 10 k BoW dictionary for regions-based feature encoding (extracted from late convolutional layers of deep VGG-16) followed up with their cross-matching degrades the real-time performance. Secondly, employment of object-centric deep VGG-16 model results in a system attempting to put more emphasis on objects rather than the place itself. This reflects on the regions-based pooled feature and leads to failure cases. Also, the regional approach proposed in [7] hinders the identification of individual static place-centric regions that can be more effective under condition and viewpoint variations.

To bridge those research gaps, this article proposes a holistic approach targeted for a CNN architecture comprising a small number of layers pretrained on a scene-centric [18] image database to reduce the memory and computational costs. The proposed method detects salient CNN-based regional features and combines them with vector of locally aggregated descriptor (VLAD) [19] adapted specifically for the VPR problem. The motivation behind employing VLAD comes from its better performance in various CNN-based image retrieval tasks utilizing a smaller visual word dictionary [19], [20] compared to BoW [11]. To the best of our knowledge, this is the first work that combines novel lightweight CNN-based regional features with VLAD encoding adapted for VPR.

As opposed to [7] which uses object-centric VGG-16 architecture and employs a cross-convolution based regional extraction approach (resembles [15]), the proposed VPR technique here, is different both in identification and extraction of regional features (discussed in detail, in Section III-B). The approach presented in this article showcases enhanced accuracy by employing middle convolutional layer of the

Manuscript received May 20, 2019; revised October 19, 2019; accepted November 12, 2019. Date of publication December 27, 2019; date of current version April 2, 2020. This paper was recommended for publication by Associate Editor J. Kober and Editor F. Chaumette upon evaluation of the reviewers' comments. This work was supported in part by the UK Engineering and Physical Sciences Research Council (EPSRC) through Grant EP/R02572X/1 and Grant EP/P017487/1 and in part by the RICE Project funded by the National Centre for Nuclear Robotics (NCNR) Flexible Partnership Fund. *(Corresponding author: Ahmad Khaliq.)*

A. Khaliq, S. Ehsan, and K. McDonald-Maier are with the Embedded and Intelligent System Laboratory, School of Computer Science and Electronic Engineering, University of Essex, Colchester CO4 3SQ, U.K. (e-mail: ahmad.khaliq@essex.ac.uk; sehsan@essex.ac.uk; kdm@essex.ac.uk).

Z. Chen is with the Vision for Robotics Laboratory, ETH Zurich, Zurich 8092, Switzerland (e-mail: chenze@ethz.ch).

M. Milford is with the Australian Centre for Robotic Vision and School of Electrical Engineering and Computer Science, Queensland University of Technology, Brisbane, QLD 4000, Australia (e-mail: michael.milford@qut.edu.au).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TRO.2019.2956352

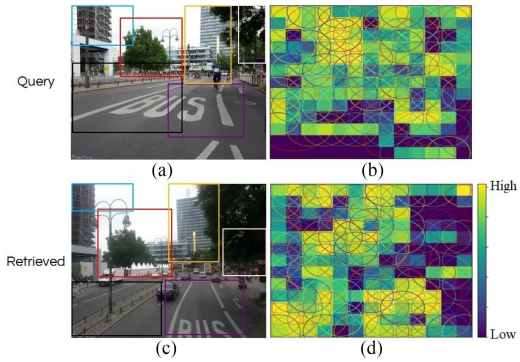

Fig. 1. For a query image (a), the proposed Region-VLAD approach successfully retrieves the correct image (c) from a stored image database under significant viewpoint- and condition-variation. (b) and (d) represent their CNN based meaningful regions identified by our proposed methodology.

eight-layered CNN architecture. Evaluation on several viewpoint- and condition-variant benchmark place recognition datasets shows an average performance boost of 13% over state-of-the-art VPR algorithms in-terms of AUC computed under precision-recall curves. Fig. 1 shows that for a query image (a), our proposed system retrieved image (c) from the stored database. (b) and (d) highlight the salient regions which our proposed methodology identified under strong visual changes.

The rest of this article is organized as follows. Section II provides the related work for VPR and other image retrieval tasks. In Section III, the proposed methodology is presented in detail. Section IV illustrates the implementation details and performance evaluation of the proposed VPR framework on several benchmark datasets. Section V concludes this article.

II. Related Work

This section provides an overview of major developments in VPR under simultaneous viewpoint and appearance changes using handcrafted and CNN-based features. Other image retrieval tasks with their feature extracting and encoding approaches are further discussed and differentiated from VPR based image retrieval tasks.

FAB-MAP [13] is the first work that used handcrafted SURF feature descriptors combined with BoW encoding for VPR. It demonstrated robustness under viewpoint changes by taking advantage of the invariance properties of SURF. Another sequence-based image matching technique, SEQSLAM [14] has shown remarkable performance under severe appearance changes. However, it is unable to deal with simultaneous condition and viewpoint variation.

The first CNN-based VPR system is introduced in [5], which is followed in [6], [21], and [22]. Chen *et al.* [5] used Overfeat [23] trained on ImageNet. Eynsham [13] and QUT datasets with multiple traverses of the same route exhibiting environmental changes are used for benchmarking. Using the Euclidean distance on the pooled layers' responses, test images are matched against the reference images. Sünderhauf *et al.* [6] and Panphattarasap and Calway [22] used landmark-based approaches coupled with the pretrained CNN models. Chen *et al.* [24] introduced two CNN models for the specific task of VPR (named AMOSNet and HybridNet), which trained and fine-tuned the object-centric CaffeNet [10] on a 2.5 million Specific PlacEs Dataset (SPED). The place-recognition centric SPED consists of thousands of places with severe-condition variance among the same places over different times of the year. The results showed that with spatial pyramidal pooling (SPP) employed on middle and late convolutional layers, HybridNet outperformed AMOSNet, CaffeNet, and PlaceNet on four publicly available datasets exhibiting strong appearance and moderate viewpoint changes [24].

Chen *et al.* [7] presented a VPR approach that identifies pivotal landmarks by directly extracting prominent patterns based on responses of late convolutional layers of deep object-centric VGG-16 model. Recently, Chen *et al.* [8] introduced a context-flexible attention model and combined it with a pretrained object-centric VGG-16 fine-tuned on SPED [24] to learn more powerful condition-invariant regional features. This system has shown state-of-the-art performance on severe condition-variant datasets. However, the efficiency of the framework may be compromised if there is a simultaneous strong viewpoint and condition variations. Moreover, performance and efficient resource usage remain two important aspects to be looked upon in real-life robotic VPR applications.

Image retrieval tasks which either rely on handcrafted features, such as, local SIFT and SURF features [3], [4] or combining these with convolutional and fully connected layers of deep/shallow CNNs [2], [5], [25], BoW, or support vector machine (SVM) [26] are employed for classification, detection, and recognition [15], [17] purposes. As an alternative for BoW feature encoding scheme, several other approaches including Fisher vector [27] and VLAD have shown promising results with smaller visual words vocabularies [19]. To perform instance level image retrieval where objects from the same category are to be separated, Yue-Hei Ng *et al.* [25] suggested to combine rich spatial middle convolutional layers' features with VLAD encoding. Kim *et al.* [28] have used MSER [29] for regions identification, coupled with SIFT feature description within the identified regions and described each region/bundle as a fix sized VLAD, named as PBVLAD. Two dimensional (2-D)-based localization methods generally offer efficient database management at low accuracy cost, whereas 3-D-based techniques are computationally complex but more reliable in localization. Sattler *et al.* [30] refute this notation by combining 2-D-based approaches with SfM-based postprocessing and shown better performances then structure-based methods. However, such postprocessing takes significant longer run-times which is out of scope of this article since our proposed VPR system works like a 2-D-based framework with an aim to improve the retrieval performance while reducing the computation complexities.

The advent of several feature pooling techniques including Sum-Pooling [16], Max-Pooling [17], Spatial Max-Pooling [31], and Cross-Pooling [15] employed in deep CNNs have demonstrated performance boost in image classification/recognition and object detection [15], [17] tasks. All these pooling methods process the convolutional layers' feature maps as a whole to pick prominent patterns, and images containing fewer objects make feature maps sparse in nature and finding single region of interest becomes relatively easier for object/image recognition and classification. However, such image retrieval tasks are different in nature from the VPR systems where recognizing a place which undergoes diverse changes due to illumination, winter–summer transition, or viewpoint variance added by different capturing angles is quite challenging because appearance of the place changes and makes it difficult to identify the common regions. For VPR, when such external tasks based pretrained CNNs [10] are integrated with the abovementioned feature pooling techniques, the convolutional layers feature maps focus on the trained objects such as vehicles, pedestrians, and other time varying objects which are not suitable for place recognition [7]. Therefore, it is still questionable for a generic VPR system to efficiently deal with simultaneous viewpoint and condition variations when employing CNN-based local features pretrained on other image retrieval tasks.

Recently, Teichmann *et al.* [32] trained the landmark detectors [17], [33] with the newly introduced 1.2 M Google Landmark dataset containing manually annotated 15 k landmark categories (such as, buildings, monuments, and bridges). Observing that not all the visual words get associated with the feature descriptors results into many zero regional residuals, their proposed R-VLAD technique overcomes it by normalizing the regional residuals [32]. Precisely, it down-weights all the regional residuals and stores a single aggregated regional descriptor per image. Custom landmark detectors including ASMK [34], RMACB [33], RMAC [17], and selective search [35] are incorporated for the regional search and coupled with R-VLAD on deep CNNs. We can expect further boost in our proposed VPR framework with

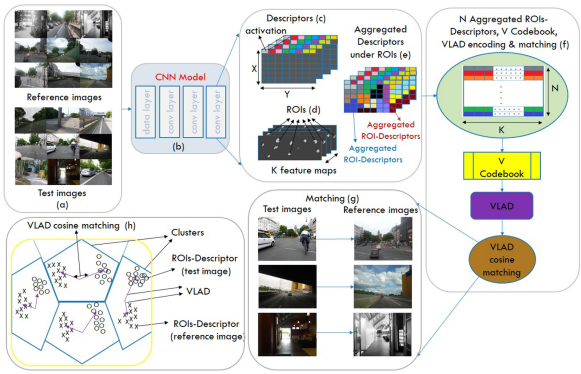

Fig. 2. Workflow of the proposed VPR framework is shown here. Test/reference images are fed into the CNN model, Region-of-Interests (ROIs) are identified across all the feature maps of the convolution layer and their compact VLAD representation is stored for image matching.

the integration of R-VLAD [32]. Chen *et al.* [8] have shown that the state-of-the-art regions-based image retrieval techniques including attentive attention [36] and fixed context [37] are not generally efficient for VPR under strong visual changes.

III. Method

In this section, the key steps of the proposed methodology are described in detail. It starts by stacking activations of feature maps for retrieving local descriptors, followed up with the identification of distinguishing regional patterns. It then illustrates the aggregation of local feature descriptors lying under those identified salient regions. Finally, it shows how to retrieve the compact VLAD representation using the extracted CNN-based regional features, later used for determining a match between two images. The workflow of the proposed methodology is shown in Fig. 2.

A. Stacking of Convolutional Activations for Making Descriptors

Given an image I as an input to the CNN model, at a certain convolutional layer, the output is a 3-D tensor M of $X \times Y \times K$ dimensions. K denotes the number of feature maps, X and Y represent the width and height of feature map/channel, respectively. We can also interpret it as M^k being a set of $X \times Y$ activations/responses for kth feature map where $k = \{1, 2, \ldots, K\}$. For K feature maps in the convolution layer, we stack each activation at some certain spatial location into K-dimensional local feature as shown with different colors in Fig. 2(c). D^L in (1) represents the K-dimensional d_l feature descriptor(s) at Lth convolutional layer of m_c model.

$$D^L = \{d_l \in M^K \;\; \forall\, l \in \{(i,j) \mid i = 1, \ldots, X; j = 1, \ldots, Y\}\}, \quad L \in m_c. \tag{1}$$

B. Identification of Regions of Interest

To extract region-based CNN features, the most prominent regions need to be identified. Two or more activations are considered to be connected and represented as a region if they are neighbors and have approximately the same value. For K feature maps, each region is denoted by $G_h, \forall\, h \in \{1, \ldots, H\}$ where H is the total identified regions at Lth convolution layer, visualized in Fig. 2(d)/Fig. 4.

The mean energy of each G_h region is calculated by averaging all a_h activations lying under the region. In (2), a_h^f represents the fth activation lying under the G_h region and E^L denotes the calculated mean energies of H regions. Based on the sorted E^L energies, top N energetic ROIs (with their bounding boxes) are picked in (3), denoted as R^L novel regions at Lth convolution layer.

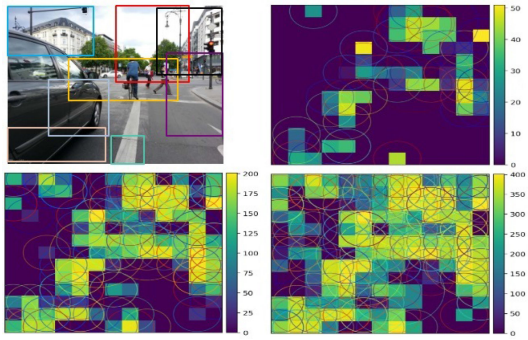

Fig. 3. Sample images of top $N = \{50, 200, 400\}$ ROIs identified by our proposed approach at Lth convolutional layer; CNN-based identified regions put emphasis on static objects such as, buildings, trees, and road signals.

Fig. 3 illustrates the top $N = \{50, 200, 400\}$ novel R^L regions identified by our proposed regions-based VPR system. Our novel CNN-based identified regions strongly concentrate on the static objects including buildings, trees, and road signals. D^L local descriptors in (1) which fall under the bounding boxes of R^L regions in (3), aggregated in (4) to retrieve CNN-based regional features. Intuitively, each regional feature is $1 \times K$ dimensional f_t vector where q be the R_t^L region under which D_q^L descriptors fall. For N novel regions, (5) represents $N \times K$ dimensional F^L region-based CNN features representing an image at Lth convolutional layer [visually shown in Fig. 2(e) and Fig. 2(f)]

$$E^L = \left\{ \frac{1}{|G_h|} \sum_f a_h^f, \; \forall\, a_h^f \in G_h \right\} \tag{2}$$

$$R^L = \{G_t \; \forall\, t \in \{1, \ldots, N\}\} \tag{3}$$

$$f_t = \sum_{q \in R_t^L} D_q^L \tag{4}$$

$$F^L = \{f_t \; \forall\, t \in \{1, \ldots, N\}\}. \tag{5}$$

In comparison, Chen *et al.* [7] first identified regions, calculated their mean energies, and selected $N = 200$ energetic regions. Precisely, N regional activations at Lth convolution layer were mapped onto the L–1th convolutional feature maps, and aggregation of modified cross-mapped regions-based local descriptors at L–1th convolution layer was carried out for feature extraction. Note that depending upon the quantity of activations per ROI(s) at Lth convolution layer and receptive field of the filter (e.g., 3×3, 5×5) for cross mapping of Lth convolution layer regions at L–1th layer, the bounding box (area) per cross-mapped regional feature varies for the work done in [7].

Furthermore, Fig. 4 illustrates that the identified ROIs from two feature maps (M^1 and M^2) at Lth convolutional layer with Region-VLAD and Cross-Region-BoW [7] are different in quantity and size/activations per region(s). Thus, the computed regional mean energies of [7] are different from the mean energies of regions identified by our approach. Our approach identifies 36 and 40 ROIs from feature map M^1 and M^2, shown with different colors. Later, based on their computed mean energies, top N energetic regions are selected from H identified regions at Lth convolutional layer, shown in Fig. 3. The eight-connected component-based regional approach in Cross-Region-BoW [7] identifies six and four yellow colored ROIs for feature map M^1 and M^2. As explained above, N energetic regional feature extraction for [7] is carried out by first selecting N energetic regions at Lth layer (see Fig. 4) followed up with their mapping at L–1th convolution layer and aggregation of cross-mapped regions-based local descriptors at L–1th convolution layer (not shown in the figure). Exemplars exhibiting the identified regions by Cross-Region-BoW [7] and with our proposed Region-VLAD framework are shown in Fig. 5. We observe that regional patterns covering more areas similar to [7] hinder the identification of individual place-centric

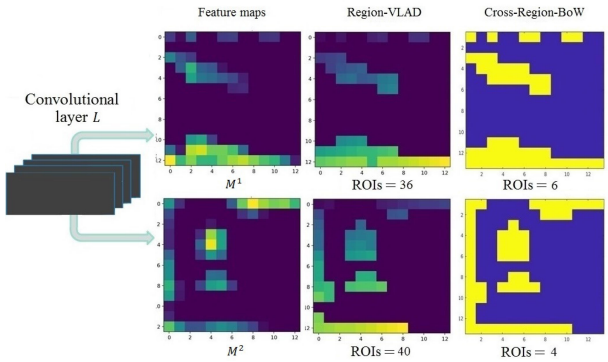

Fig. 4. Employing two features maps M^1 and M^2, sample images of ROIs identified by Region-VLAD and Cross-Region-BoW [7] are shown here. Note that feature maps (1st column) illustrate the intensities of a activations. However, regardless of the intensity, each identified G_h region per feature map for Region-VLAD (2nd column) is indicated with a different color, i.e., 36 and 40 colored regions for feature map M^1 and M^2. For Cross-Region-BoW (3rd column), all the regions are denoted as yellow patterns, i.e., 6 and 4 ROIs for M^1 and M^2 feature maps.

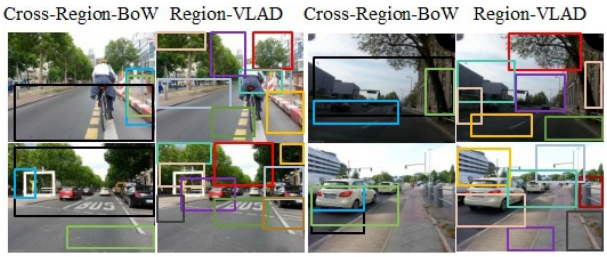

Fig. 5. Sample images of ROIs identified with Cross-Region-BoW [7] and Region-VLAD are shown here. Our regional approach subdivides each image into large number of most contributing regional blocks.

instances vital in recognizing places under changing conditions and viewpoints.

C. Regional Vocabulary and Extraction of VLAD for Image Matching

VLAD adopts K-means [11] based vector quantization, accumulates the residues of features quantized to each dictionary cluster, and concatenates those accumulated vectors into a single feature representation. A separate dataset of 2.6 k images is collected and afore-described regions-based feature extraction is employed for generating a regional vocabulary. To learn a diverse vocabulary, we employed 1125 place-recognition centric images of 365 places from Query247 [38] (taken at day, evening, and night times). Other images include a benchmark place recognition dataset St. Lucia [24] with 1 k frames of two traverses captured in suburban environment at multiple times of the day. The left over images consist of multiple viewpoint- and condition-variant traverses of urban and suburban routes collected from *Mapillary* (previously employed by [6] and [7] for capturing place recognition datasets). K-means is employed for clustering the $2600 \times N \times K$ dimensional regional features into V regions such that o_u in (6) represents the uth region of C^L codebook

$$C^L = \{o_u \; \forall \; u \in \{1, \ldots, V\}\}, \; V \in \{64, 128, 256\}. \tag{6}$$

Using the learned codebook, F^L regions of benchmark test / reference traverses are quantized in (7) to predict the clusters or labels Z^L, where α is the quantization function. Employing regions-based F^L feature, predicted labels Z^L and regional codebook C^L, summed residue v corresponding to each uth region can be retrieved using (8)

$$Z^L = \alpha(F^L). \tag{7}$$

In (8), for all the F^L regional features that fall in uth region of the C^L codebook, the residues of F^L_u regions and C^L_u codebook's region center are summed. Sometimes, few regions/words appear more frequently in an image than the statistical expectation known as visual word burstiness [39]. Standard techniques include power normalization [40] is performed in (9) to avoid it where each $1 \times K$ dimensional residue v_u undergoes nonlinear transformation γ. In (10), power normalization is followed by l_2 normalization. For each image, l_2 normalized residues corresponding to V regions are stored in (11) to get final $V \times K$ dimensional VLAD representation S^L.

To match a test image "A" against the reference image "B" in (12), the dot/scalar product of their uth regional VLAD components $S^{L^A}_u$ and $S^{L^B}_u$, each with dimension $1 \times K$ reaches to an individual regional matching score $j^{A,B}_u$, as visualized in Fig. 2(h)

$$v_u = \sum_{F^L_u : Z^L_u = C^L_u} F^L_u - C^L_u \tag{8}$$

$$v_u := sign(v_u) \|v_u\|^\gamma \tag{9}$$

$$v_u := \frac{v_u}{\sqrt{v_u^T v_u}} \tag{10}$$

$$S^L = \{v_u \; \forall \; u \in \{1, \ldots, V\}\} \tag{11}$$

$$j^{A,B}_u = \frac{\left(S^{L^A}_u\right) \cdot \left(S^{L^B}_u\right)}{\| \left(S^{L^A}_u\right) \| \| \left(S^{L^B}_u\right) \|}. \tag{12}$$

All the scalar $j^{A,B}_u$ scores for V regions are summed up in (13) to get final single $J^{A,B}$ matching score. For each test image "A," the cosine matching in (12) is performed against all the reference images and finally, reference image "X" with the highest similarity score is picked as a matched image using (14)

$$J^{A,B} = \sum_{u=1}^{V} j^{A,B}_u \tag{13}$$

$$P^A = \arg\max_X J^{A,X}. \tag{14}$$

IV. Datasets, Implementation Details, Results, and Analysis

This section presents the implementation details of our proposed system which will attempt to evaluate its run-time performance for real-time robotic VPR applications. Comparison of the proposed method with state-of-the-art VPR and image retrieval algorithms has been conducted over several benchmark datasets and the obtained results are stated. The section ends by displaying the results on correctly matched and mismatched scenarios of our proposed Region-VLAD framework alongside a discussion.

A. Benchmark Place Recognition Datasets

More specifically, challenging benchmark VPR datasets *Berlin A100*, *Berlin Halenseestrasse*, and *Berlin Kudamm* (see [7] for detailed introduction), collected from crowd-sourced geotagged photo-mapping platform *Mapillary* are used to evaluate the proposed VPR framework. Each dataset covers two traverses of the same route uploaded by different users. One traverse is used as R reference traverse and the other traverse is employed as T test traverse (see Table I). R' represents the reduced reference traverse which matches with T' test traverse (discussed in Section IV-E). Another dataset, *Gardens Point* was captured at QUT campus with one traverse taken in daytime on left side walk and the other traverse was recorded in right side walk at night time [24]. The *Synthesized Nordland* dataset was recorded on a train journey with one traverse taken in winter and the other traverse was recorded in spring. Viewpoint variance was added by cropping frames of summer traverse to keep 75% resemblance [8]. For *Berlin A100*, *Berlin Halenseestrasse*, and *Berlin Kudamm*, geotagged information is used for ground truth with 0 to ± 2 frame tolerance. For *Gardens Point*

TABLE I
BENCHMARK PLACE RECOGNITION DATASETS

Dataset	Environment	Variation		T	R	T'	R'
		Viewpoint	Condition				
Berlin A100	urban	moderate	moderate	81	85	70	64
Berlin Haleenseetrasse	urban, suburban	very strong	moderate	67	157	50	138
Berlin Kudamm	urban	very strong	moderate	222	201	166	151
Gardens Point	campus	strong	strong	200	200	152	150
Synthesized Nordland	train	moderate	very strong	1622	1622	1221	1217

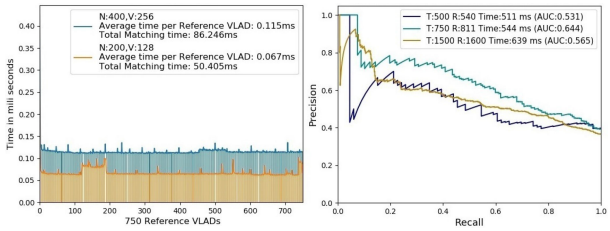

Fig. 6. Left: Matching times for 1 test VLAD against 750 reference VLADs are presented. Right: AUC-PR performance and retrieval time of Region-VLAD are reported while adding more images in T test and R reference traverses.

and *Synthesized Nordland*, the ground truth data is obtained by parsing the frames and maintaining place level resemblance with 0 to ± 3 and 0 to ± 2 frame tolerance.

B. Setup, Implementation Details, and Scalability

The proposed VPR system is implemented in Python 3.6.4 framework and the system average runtime over 5 iterations is recorded with 1125 images. AlexNet pretrained on Places365 dataset is employed as a CNN model for region-based features extraction with 256×256 input image size. For all the baseline experiments, we utilize middle *conv3* convolutional layer only due to its better performance in various VPR approaches [6], [22].

For a single image, a forward pass takes around an average 0.305 ms using Caffe on NVIDIA P100 and 15.57 ms employing Intel Xeon Gold 6134 @3.2 GHz. We extract N ROIs with total time on par with state of the art methods [7] (see Table II). The VLAD representations are retrieved and matched using N ROIs mapped on V clustered dictionary C^L (trained using N ROIs per image of 2.6 k dataset). For direct comparison with [7], we use $N = 200$ with $V = 128$. The results are also reported for $N = 400$ with $V = 256$. Table II shows that for $N = \{200, 400\}$ regional settings, our average VLAD matching times are $100x$ and $58x$ faster than [7].

In real-time robotic vision applications which include robotic agricultural devices, autonomous infrastructure, environmental monitoring equipment or other agriculture based use-cases, with exploration of new places, the size of the database can grow unbounded and scalability becomes an important factor to be considered [41]. Under both the regional settings, employing GPU for forward pass and CPUs for both feature extraction and VLAD encoding, the overall times for retrieving a single query VLAD are 396 and 447 ms. Whereas, Titan X Pascal GPU in [7] takes 408 ms for feature encoding per query. Fig. 6 (left) further confirms that the proposed system consumes an average 0.07 ($N = 200$) and 0.12 ms ($N = 400$) for matching VLAD representations of a single query and reference image. Therefore, the total retrieval times per query against $R = 750$ reference images are approximately around 446.405 and 533.245 ms. In comparison, Cross-Region-BoW [7] takes 7 ms for matching features of one test and one reference image. The overall retrieval time against $R = 750$ reference images is 5.658 s which is $12x$ and $11x$ more than our proposed approaches and practically inappropriate for real-time applications. Our Region-VLAD VPR technique can store the encoded VLAD representations of all the reference frames, whereas Cross-Region-BoW needs to perform run-time cross matching of given query regions against all the reference frames' regions, and mutually matched regional features are picked.

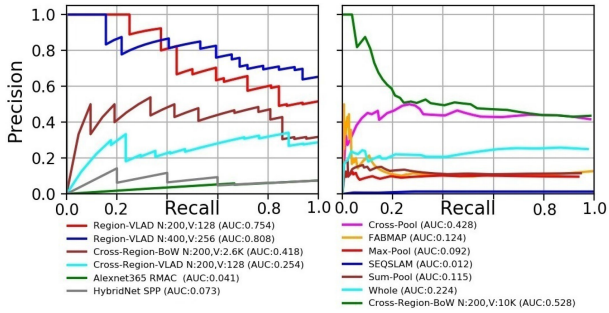

Fig. 7. AUC PR-curves for Berlin Halenseestrasse dataset are presented here. Left: PR-curves of our proposed Region-VLAD and [7] employed on AlexNet365 with VLAD and BoW encodings. Right: Comparison with state-of-the-art VPR approaches employing VGG-16.

Furthermore, Fig. 6 (right) evaluates our proposed system's run-time performance when more places are added in test and reference traverses. For each PR-curve, we employed T test and R reference images. Their VLAD representations are retrieved followed up by their cosine matching and in parallel, we record down the system's performance. We can see that as the size of test and reference traverses increases, the AUC-PR curves remain higher where "Time" represents the overall matching period for a single test image against R reference traverse. This shows that the system is capable to handle large number of reference/database images while maintaining performance both in accuracy and retrieval time. It should be noted that, Chen *et al.* [7] used MATLAB which is practically slower than Python, but we have employed CPUs in comparison to the work done in [7] which used GPU.

C. Comparison Methods

To show the dominance of our novel place-centric regions finding approach, we replaced VGG-16 with AlexNet365 in [7] and combined the regional features with VLAD and BoW encodings, named as Cross-Region-VLAD and Cross-Region-BoW. For a fair comparison, using 2.6 k dataset, we trained a separate regional vocabulary employing *conv4* for regions identification and *conv3* for feature extraction. Keeping $N = 200$, we used $V = 128$ for Cross-Region-VLAD and $V =$ 2.6 k for Cross-Region-BoW. Furthermore, results are also reported for HybridNet with SPP [24] employed on *conv5* of the model. We also integrated RMAC [17] on AlexNet365 while performing power- and l2-normalization on the retrieved regional features. Similar to the work done in [7], mutual regions are filtered using cross matching, their scores are summed up and maximum matching score is considered for retrieval.

PR-curves across all other image retrieval approaches including Cross-Pool, Max-Pool, Sum-Pool, Whole and state-of-the-art VPR approaches FABMAP and SEQSLAM are taken from [7]. Chen *et al.* [7] employed *conv5_2* of deep object centric VGG-16 as features representation. However, Cross-Region-BoW [7] with deep VGG-16 model used *conv5_3* for landmarks identification and *conv5_2* for feature extraction. Standard FABMAP implementation and three sequential frames configuration for SEQSLAM were used in [7].

D. Precision Recall Characteristics

In image retrieval tasks where there is a moderate to large class imbalance which means the positive class samples are quite rare as compared to the negative classes, precision-recall curves are usually employed as evaluation metric [12]. For all the benchmark datasets, we first calculate the difference in AUC-PR performance of [7] and Region-VLAD, determine their average which comes around an overall 13% performance improvement.

1) Berlin Halenseestrasse: In Fig. 7 (left), the proposed Region-VLAD PR-curves for *Berlin Halenseestrasse* dataset significantly

TABLE II
COMPARISON OF OUR PROPOSED METHOD WITH CROSS-REGION-BOW [7]

Methodology		Our Region-VLAD																	Cross-Region-BoW [7]
Model		AlexNet365																	VGG16
Images		1125																	1000
GPU/CPU		NVIDIA P100	Intel(R) Xeon(R) Gold 6134 CPU @ 3.20GHz with 32 cores, 64GB RAM																Titan X Pascal GPU
Forward pass time (ms)		0.305	15.574																59
ROIs "N"		50			100			200			300			400			500			200
Extraction time (s)		0.328			0.361			0.394			0.402			0.443			0.452			0.349
Regions "V"		64	128	256	64	128	256	64	128	256	64	128	256	64	128	256	64	128	256	10k Visual words
Matching time (ms)	**VLAD encoding**	1.33	2.05	3.58	1.55	2.28	3.79	1.91	2.4	4.03	1.99	2.68	4.28	2.13	2.96	4.54	2.36	3.16	4.75	7
	VLAD matching	0.05	0.06	0.12	0.05	0.08	0.13	0.05	0.07	0.12	0.04	0.07	0.12	0.05	0.08	0.12	0.05	0.07	0.12	

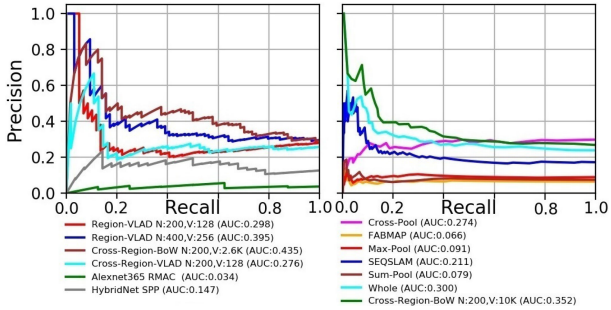

Fig. 8. AUC PR-curves for Berlin Kudamm dataset are presented here. Left: PR-curves of our proposed Region-VLAD and [7] employed on AlexNet365 with VLAD and BoW encodings. Right: Comparison with state-of-the-art VPR approaches employing VGG-16.

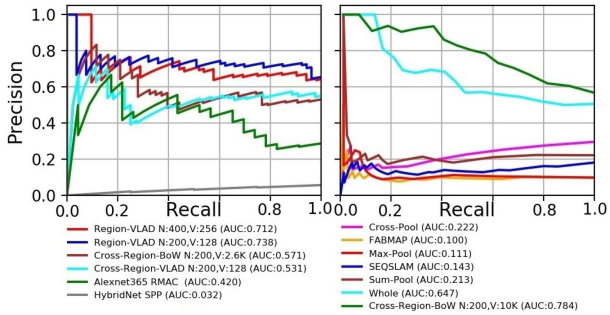

Fig. 9. AUC PR-curves for Berlin A100 dataset are presented here. Left: PR-curves of our proposed Region-VLAD and [7] employed on AlexNet365 with VLAD and BoW encodings. Right: Comparison with state-of-the-art VPR approaches employing VGG-16.

outperforms all other state-of-the-art methods. Surprisingly, Cross-Region-VLAD PR-curve underperformed with a big margin. This mimics that the better AUC-PR performance of our proposed approach is encouraged with the use of our novel regional features. Furthermore, investigations on Cross-Region-VLAD suggest that under strong viewpoint changes, the mapping of cross-convoluted regional patterns [7] over the vocabulary for VLAD retrieval, results in nonuniform feature distribution. Although, normalization is carried out, many zero regional residues exist in the VLAD representation which reflects on the PR-curves. Cross-Region-BoW only considers the mutually matched regions and exhibits better results. Moreover, RMAC which is state-of-the-art in other image retrieval techniques and SPP, both are sensitive under strong viewpoint variation, thus under-performed on this dataset.

Although, FABMAP is robust under viewpoint variation but still underperforms on this dataset just like SEQSLAM, a whole image-based technique which subtracts patch-normalized sequence of frames. Cross-Pool employs a similar idea of pooling as Cross-Region-BoW, so both have achieved nearly the similar PR-curves, whereas other pooling techniques underperformed. It is worth noting that even with smaller regional dictionaries, our proposed Region-VLAD framework still achieves better results than VGG-16 based Cross-Region-BoW [7] and other methodologies. It highlights the potential of our shallow CNN based regional features robustness under strong viewpoint variations.

2) Berlin Kudamm: Due to urban environment, too many dynamic and confusing objects such as vehicles, trees, and pedestrians with homogeneous scenes lead to perceptual aliasing coupled with severe viewpoint changes makes it a challenging dataset. Fig. 8 (left) shows that our proposed Region-VLAD approach still manages to achieve better results. AlexNet365 combined with Cross-Region-BoW claims state-of-the-art results with $V = 2.6$ k regional vocabulary. RMAC and SPP again underperformed. This is apparently because VPR is different from other image retrieval and recognition systems where a single object majorly covers the whole image. Therefore, Sum-Pool, Max-Pool, and RMAC which perform relatively well in such vision-based tasks, did not performed well in VPR under strong viewpoint and appearance changes.

In Fig. 8 (right), due to resemblance among the places captured in sequence, Whole and SeqSLAM with their whole-image based approach have shown better performance. With higher precision at start and as recall increases, Region-VLAD PR-curves are quite similar but covering more AUC than Whole, SeqSLAM, Cross-Pool, and VGG-16 Cross-Region-BoW.

3) Berlin A100: This dataset exhibits moderate viewpoint and moderate conditional changes coupled with dynamic objects. PR-curves are displayed in Fig. 9. It is quite evident that our Region-VLAD approach in Fig. 9 (left) achieves similar results as state-of-the-art VGG-16 Cross-Region-BoW [7]. AlexNet365 combined with cross-regional approach of [7] achieves similar and better results for BoW and VLAD. SPP employed on HyridNet was found not very convincing. It might be because HybridNet is fine-tuned on SPED which contains minimal dynamic instances among the same place(s) captured over multiple times of the year.

Against our approach, RMAC on AlexNet365 achieves comparable and better performance than FABMAP and pooling techniques including Sum-Pool, Max-Pool, and Cross-Pool. Since condition and viewpoint variations are not strong in this dataset, RMAC and other approaches have also shown better performance. A deep analysis on the dataset reveals varied time intervals between the captured frames due to which SEQSLAM underperformed on this dataset. Overall, our proposed Region-VLAD achieved second best performance after VGG-16 Cross-Region-BoW [7].

4) Synthesized Nordland: In comparison to other approaches, PR-curves in Fig. 10 (left) show that our proposed approach works relatively well on this dataset but RMAC and SPP achieve better performance. Employing deep VGG-16, Max-Pool and Sum-Pool have not shown better results and similar whole image-based techniques, i.e., SEQSLAM and Whole exhibit similar PR-curves.

Since in HybridNet, fine-tuning the CNN model with SPED induced condition invariance. Therefore, employing SPP on HybridNet has shown superior performance on this dataset (exhibiting strong conditional changes). In comparison, scene-centric AlexNet365 integrated with Cross-Region-BoW and Cross-Region-VLAD outperformed deep ImageNet-centric VGG-16 based Cross-Region-BoW [7]. This highlights the importance of CNN training.

5) Gardens Point: Both the *Gardens Point* traverses exhibit stronger lightning variations with adequate temporal coherence between the frames. Fig. 11 shows that our Region-VLAD approach

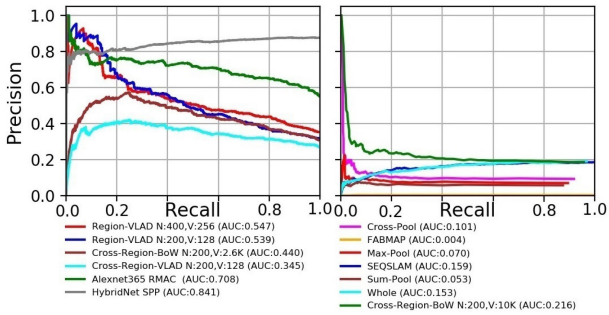

Fig. 10. AUC PR-curves for Synthesized Nordland dataset are presented here. Left: PR-curves of our proposed Region-VLAD and [7] employed on AlexNet365 with VLAD and BoW encodings. Right: Comparison with state-of-the-art VPR approaches employing VGG-16.

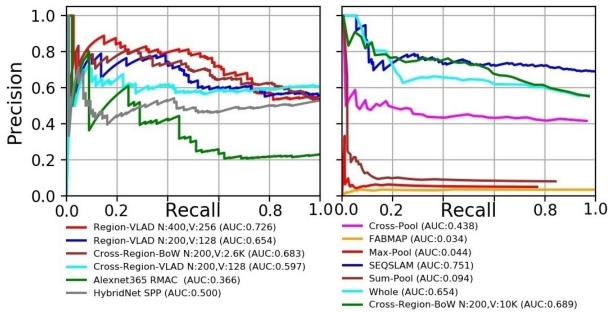

Fig. 11. AUC PR-curves for Gardens Point dataset are shown here. Left: PR-curves of our proposed Region-VLAD and [7] employed on AlexNet365 with VLAD and BoW encodings. Right: Comparison with state-of-the-art VPR approaches.

achieves similar and better performance than Cross-Region-BoW, Cross-Region-VLAD, Whole, RMAC, and SPP. Taking advantage from the sequential information, SEQSLAM has shown state-of-the-art performance. Cross-Region-BoW and Cross-Region-VLAD integrated with AlexNet and VGG-16 exhibit similar performances but approaches including Sum-Pool, Max-Pool, and FABMAP relatively underperformed.

E. Matching Score Thresholding

By nature, PR curves do not consider True Negative cases (correctly missed the nonexisting events/classes) [12]. So, in order to tackle such tricky situations, we employ T test traverse and R' reference traverse from all the datasets so that $T - T'$ queries can be treated as new places (see Table I). Fig. 12 visualizes the results of the proposed Region-VLAD framework before (left column) and after (right column) the match score thresholding. On the basis of matching scores, y-axis differentiates the True Positive (TP), False Negative (FN), False Positive (FP), and True Negative (TN) events shown with different colored curves, where length of the curves in x-axis denotes the number of images which the events contain. The threshold is an average of TN scores of R' reference traverses of the benchmark datasets. Due to limited space, results are reported for two datasets only. Upon thresholding in Fig. 12 (right column), Region-VLAD for *Berlin Halenseestrasse* dataset missed $FN = 2$ correctly matched images and successfully filtered 10 queries out of $TN = 17$. The same behavior is observed for *Berlin A100* dataset. In scenarios when the system comes across previously observed places as well as new places, it becomes increasingly challenging to successfully retrieve the correct matches (TPs), discard incorrect matches (TNs), while reducing FPs (retrieved incorrect matches) and FNs (discarded correctly retrieved matches). It is evident that Region-VLAD not only boosts up the AUC under PR-curves but also deals efficiently in assigning low scores to TN queries (green curves).

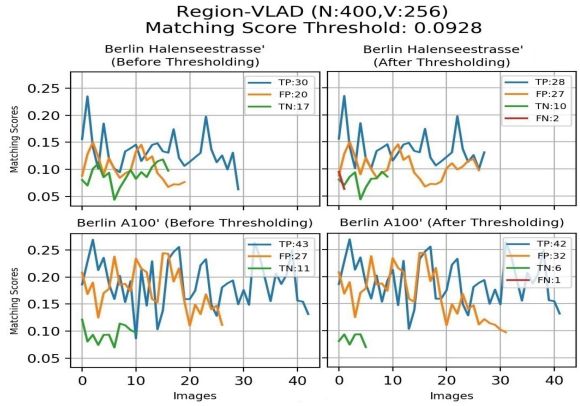

Fig. 12. Left column presents graphs for *Berlin Halenseestrasse* and *Berlin A100* before thresholding and right column graphs showcase the change in TP, FP, TN, and FN upon thresholding. Our proposed Region-VLAD framework assigned low score to the T-T' or TN queries.

Fig. 13. Sample correctly retrieved matches using the proposed VPR framework are presented here; it identifies common regions across the queries and retrieved images under strong viewpoint and appearance variations.

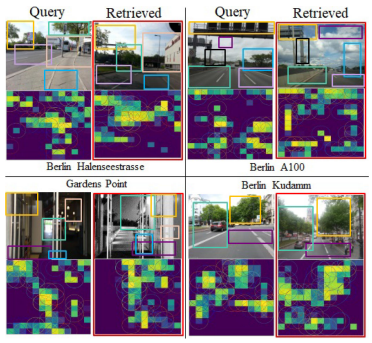

Fig. 14. Sample incorrectly retrieved matches using the proposed VPR framework are presented here; each query and the retrieved database images are geographically different but exhibiting similar scenes and conditions.

F. Analysis

Figs. 13 and 14 illustrate some of the matched and mismatched scenarios. For the correct matches, taking advantage from CNN's scene-centric training, Region-VLAD identifies the common regions shown with different colored boxes under simultaneous viewpoint and appearance changes. For the mismatched scenarios, the identified top novel regions with colored boxes (trees, lamp posts) show the areas where the system is interested in and matches the scenes but wrongly recognizes the places. We have seen that Cross-Region-BoW [7] when integrated with AlexNet365 showed comparable performance but at

high time computation cost. However, our Region-VLAD still outperformed Cross-Region-BoW [7] with smaller dictionary and low retrieval time. Also, cross-regional approach in [7] when combined with the VLAD shown inferior results which confirms the performance boost in Region-VLAD encouraged with our novel regional approach. Datasets with their evaluated results are placed at [42] and code will made available upon publication.

V. Conclusion

For visual place recognition (VPR) on resource-constrained mobile robots, achieving state-of-the-art performance with lightweight CNN architectures was highly desirable but remains a challenging problem. This article took a step in this direction and presented a holistic approach targeted for a CNN architecture comprising a small number of layers pretrained on a scene-centric image database to reduce the memory and computational cost. The proposed framework detected novel CNN-based regional features and combined them with the VLAD encoding methodology adapted specifically for computation-efficient and environment Invariant-VPR problem. The proposed method achieves state-of-the-art AUC-PR curves on significant viewpoint and condition variant place recognition datasets.

In future, it would be useful to analyze the performance of the proposed framework on other shallow/deep CNN models individually trained/fine-tuned on place recognition-centric datasets. Furthermore, instead of employing defined number of novel regions, it would be interesting to investigate the dynamic regional features selection at runtime and their performances on multiple regional vocabularies.

References

[1] S. Lowry *et al.*, "Visual place recognition: A survey," *IEEE Trans. Robot.*, vol. 32, no. 1, pp. 1–19, Feb. 2016.

[2] L. Zheng, Y. Yang, and Q. Tian, "SIFT meets CNN: A decade survey of instance retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 5, pp. 1224–1244, May 2018.

[3] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," in *Proc. Eur. Conf. Comput. Vis.*, 2006, pp. 404–417.

[4] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[5] Z. Chen, O. Lam, A. Jacobson, and M. Milford, "Convolutional neural network-based place recognition," in *Proc. Australas. Conf. Robot. Autom.*, 2014.

[6] N. Sünderhauf *et al.*, "Place recognition with convnet landmarks: Viewpoint-robust, condition-robust, training-free," in *Proc. Robot. Sci. Syst. Conf.*, 2015.

[7] Z. Chen, F. Maffra, I. Sa, and M. Chli, "Only look once, mining distinctive landmarks from convnet for visual place recognition," in *Proc. IEEE Int. Conf. Intell. Robot. Syst.*, 2017, pp. 9–16.

[8] Z. Chen, L. Liu, I. Sa, Z. Ge, and M. Chli, "Learning context flexible attention model for long-term visual place recognition," *IEEE Robot. Autom. Lett.*, vol. 3, no. 4, pp. 4015–4022, Oct. 2018.

[9] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2015.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[11] Sivic and Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2003, vol. 2, pp. 1470–1477.

[12] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.

[13] M. Cummins and P. Newman, "FAB-MAP: Probabilistic localization and mapping in the space of appearance," *Int. J. Robot. Res.*, vol. 27, no. 6, pp. 647–665, 2008.

[14] M. J. Milford and G. F. Wyeth, "SeqSLAM: Visual route-based navigation for sunny summer days and stormy winter nights," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2012, pp. 1643–1649.

[15] L. Liu, C. Shen, and A. van den Hengel, "Cross-convolutional-layer pooling for image recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 11, pp. 2305–2313, Nov. 2017.

[16] A. Babenko and V. Lempitsky, "Aggregating local deep features for image retrieval," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1269–1277.

[17] G. Tolias, R. Sicre, and H. Jégou, "Particular object retrieval with integral max-pooling of CNN activations," in *Proc. Int. Conf. Learn. Represent.*, 2015, *arXiv:1511.05879*.

[18] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1452–1464, Jun. 1, 2018.

[19] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in *Proc. IEEE Comput. Soc. Comput. Vis. Pattern Recognit.*, 2010, pp. 3304–3311.

[20] R. Arandjelovic and A. Zisserman, "All about VLAD," in *Proc. IEEE Comput. Soc. Comput. Vis. Pattern Recognit.*, 2013, pp. 1578–1585.

[21] N. Sünderhauf, S. Shirazi, F. Dayoub, B. Upcroft, and M. Milford, "On the performance of convnet features for place recognition," in *Proc. IEEE Int. Conf. Intell. Robot. Syst.*, 2015, pp. 4297–4304.

[22] P. Panphattarasap and A. Calway, "Visual place recognition using landmark distribution descriptors," in *Proc. Asian. Conf. Comput. Vis.*, 2016, pp. 487–502.

[23] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," in *Proc. Int. Conf. Learn. Represent.*, 2013, *arXiv:1312.6229*.

[24] Z. Chen *et al.*, "Deep learning features at scale for visual place recognition," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2017, pp. 3223–3230.

[25] J. Yue-Hei Ng, F. Yang, and L. S. Davis, "Exploiting local features from deep networks for image retrieval," in *Proc. IEEE Comput. Soc. Comput. Vis. Pattern Recognit. Workshop*, 2015, pp. 53–61.

[26] A. F. M. Agarap, "A neural network architecture combining gated recurrent unit and support vector machine for intrusion detection in network traffic data," in *Proc. 10th Int. Conf. Mach. Learn. Comput.*, 2018, pp. 26–30.

[27] J. Sánchez *et al.*, "Image classification with the fisher vector: Theory and practice," *Int. J. Comput. Vis.*, vol. 105, no. 3, pp. 222–245, 2013.

[28] H. Jin Kim, E. Dunn, and J.-M. Frahm, "Predicting good features for image geo-localization using per-bundle VLAD," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1170–1178.

[29] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide-baseline stereo from maximally stable extremal regions," *Image Vis. Comput.*, vol. 22, no. 10, pp. 761–767, 2004.

[30] T. Sattler *et al.*, "Are large-scale 3d models really necessary for accurate visual localization?" in *Proc. IEEE Comput. Soc. Comput. Vis. Pattern Recognit.*, 2017, pp. 1637–1646.

[31] M. Jaderberg *et al.*, "Spatial transformer networks," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2015, pp. 2017–2025.

[32] M. Teichmann, A. Araujo, M. Zhu, and J. Sim, "Detect-to-retrieve: Efficient regional aggregation for image search," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 5109–5118.

[33] A. S. Razavian *et al.*, "Visual instance retrieval with deep convolutional networks," *ITE Trans. Media Technol. Appl.*, vol. 4, no. 3, pp. 251–258, 2016.

[34] G. Tolias, Y. Avrithis, and H. Jégou, "Image search with selective match kernels: Aggregation across single and multiple images," *Int. J. Comput. Vis.*, vol. 116, no. 3, pp. 247–261, 2016.

[35] R. Tao, E. Gavves, C. G. M. Snoek, and A. W. M. Smeulders, "Locality in generic instance search from one example," in *Proc. IEEE Comput. Soc. Comput. Vis. Pattern Recognit.*, 2014, pp. 2091–2098.

[36] D. Yu, J. Fu, T. Mei, and Y. Rui, "Multi-level attention networks for visual question answering," in *Proc. IEEE Comput. Soc. Comput. Vis. Pattern Recognit.*, 2017, pp. 4187–4195.

[37] H. J. Kim, E. Dunn, and J.-M. Frahm, "Learned contextual feature reweighting for image geo-localization," in *Proc. IEEE Comput. Soc. Comput. Vis. Pattern Recognit.*, 2017, pp. 2136–2145.

[38] A. Torii, R. Arandjelović, J. Sivic, M. Okutomi, and T. Pajdla, "24/7 place recognition by view synthesis," in *Proc. IEEE Comput. Soc. Comput. Vis. Pattern Recognit.*, 2015, pp. 1808–1817.

[39] H. Jégou, M. Douze, and C. Schmid, "On the burstiness of visual elements," in *Proc. IEEE Comput. Soc. Comput. Vis. Pattern Recognit.*, 2009, pp. 1169–1176.

[40] T.-T. Do *et al.*, "From selective deep convolutional features to compact binary representations for image retrieval," *Trans. Multimedia. Comput. Commun. Appl.*, vol. 15, no. 2, 2019, Art. no. 43.

[41] C. Cadena *et al.*, "Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age," *IEEE Trans. Robot.*, vol. 32, no. 6, pp. 1309–1332, Dec. 2016.

[42] "Results and datasets," [Online]. Available: https://github.com/Ahmedest61/CNN-Region-VLAD-VPR/

Ahmad Khaliq received the B.Sc. degree in computer engineering from the National University of Science and Technology, Rawalpindi, Pakistan, in 2015.

He is currently a Postgraduate Research Student with the School of Computer Science and Electronic Engineering, University of Essex, Colchester, U.K. His current research interests include computer vision, deep learning, and SLAM.

Shoaib Ehsan received the B.Sc. degree in electrical engineering from the University of Engineering and Technology, Taxila, Pakistan, in 2003, and the Ph.D. degree in computing and electronic systems with a specialization in computer vision from the University of Essex, Colchester, U.K., in 2012.

He has extensive industrial and academic experience in the areas of embedded systems, embedded software design, computer vision, and image processing. His current research interests include intrusion detection for embedded systems, local feature detection and description techniques, image feature matching, and performance analysis of vision systems.

Dr. Ehsan is a recipient of the University of Essex Post Graduate Research Scholarship and the Overseas Research Student Scholarship. He is a winner of the prestigious Sullivan Doctoral Thesis Prize by the British Machine Vision Association.

Zetao Chen received the bachelor's degree in electrical engineering from the South China University of Technology, Guangzhou, China, in 2009, the master's degree in computer science from the University of Groningen, Groningen, The Netherlands, in 2012, and the Ph.D. degree in electrical engineering and computer science from the Queensland University of Technology, Brisbane, QLD, Australia, in 2016.

Since 2016, he has been a Postdoctoral Research Fellow with Vision for Robotics Lab (V4RL), ETH Zurich, Zurich, Switzerland. His research interests include robot localization, deep learning, and SLAM.

Michael J. Milford received the Bachelor of Mechanical and Space Engineering degree and the Ph.D. degree in electrical engineering from the University of Queensland, Brisbane, QLD, Australia.

He is currently an Associate Professor and Australian Research Council Future Fellow with the Queensland University of Technology (QUT), Brisbane, and a Chief Investigator for the Australian Centre of Excellence for Robotic Vision. He was a Research Fellow on the Thinking Systems Project with Queensland Brain Institute until 2010, when he became a Lecturer with QUT. He conducts interdisciplinary research into navigation across the fields of robotics, neuroscience, and computer vision.

Dr. Milford is a recipient of an inaugural Australian Research Council Discovery Early Career Researcher Award in 2012 and became a Microsoft Research Faculty Fellow in 2013.

Klaus D. McDonald-Maier received the Dipl.-Ing. degree in electrical engineering from the University of Ulm, Ulm, Germany, the M.S. degree in electrical engineering from the cole Suprieure de Chimie Physique lectronique de Lyon, Villeurbanne, France, in 1995, and the Ph.D. degree in computer science from Friedrich Schiller University, Jena, Germany, in 1999.

He was a Systems Architect on reusable microcontroller cores and modules with the Infineon Technologies AG.s Cores and Modules Division, Munich, Germany, and a Lecturer in Electronics Engineering with the University of Kent, Canterbury, U.K. In 2005, he joined the University of Essex, Colchester, U.K., where he is currently a Professor with the School of Computer Science and Electronic Engineering. His current research interests include embedded systems and system-on-a-chip design, security, development support and technology, parallel and energy efficient architectures, and the application of soft computing and image processing techniques for real-world problems.

Dr. McDonald-Maier is a member of the Verband der Elektrotechnik Elektronik Informationstechnik and the British Computer Society, and a Fellow of the Institution of Engineering and Technology.